PROCESSING NATURAL LANGUAGE TEXT

David G. Hays

October 1966

P-3461

PROCESSING NATURAL LANGUACE TEXT

David G. Hays*

The RAND Corporation, Santa Monica, California

The skills of reading and writing are among the first taught in school because the ability to process text in at least one natural language is the fundamental intellectual skill. Cultures that do not possess it reach their zenith with the making of epic verse; medical science is as far beyond them as the stars and planets. We are just learning how to write these skills into computer programs, and perhaps it is not too much to hope that when--in the remote future--our computers can really process ordinary text our intellectual prowess will reach a level as much above the one we live on today as ours is above that of the unlettered aborigines.

Let us consider briefly just two applications: publication of scientific findings, which I will call documentation, and content analysis as a tool of psychological research. We examine the design of systems first, then some prerequisites to both.

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff. This paper was prepared for a Seminar on Computational Linguistics held by the National Institutes of Health, 6-7 October 1966.

The most appropriate starting point is at the author's typewriter. If text is acquired this early, the computer can serve as editorial aide, help with typesetting for original publication, carry most of the burden of making up secondary publications, and follow through with retrospective searches, analyses of the vocabularies of disciplinary specialties, and so on into the unpredictable future.

Thre· methods of acquisition are in use, and a fourth is developing. The simplest is to use a typewriter that punches paper tape; the tape can be fed directly to a large computer. Another is to use a typewriter with a little computing power built in. The third is to couple the author's typewriter directly to a central computer, providing hardware and programs to distribute its capacity among many typewriters at work simultaneously. Finally, typewritten sheets can be fed to a machine that identifies the characters written on them and stores a digital encoding.

The most attractive of these methods is surely the time-shared on-line console. Each keystroke is captured as it is made, and the computer's power is instantly available to the author. A well-designed system of this kind provides for immediate correction of typing errors; allows abbreviations (and expands them in the final copy); and will eventually insert bibliographic references from the author's file. But all of these methods allow editing of copy once it is stored in rough draft, so that the menial task of retyping is obviated.

Like the cortex, a good page of print has its architecture, related to the functions of its parts. Rarely can a technical subject be expounded clearly in an uninterrupted flow of words. Typically, many headings and subheadings, footnotes, figures with captions, and other apparatus are needed to set off the topics treated. Each of these parts is identified, in a well-made book or journal, by distinctive type and spacing. A book designer is an artist; when he does his work well, every page is pleasing to the reader's eye. Aesthetic standards should be applied to book design, but its price is justified by utilitarian concerns.

Printing with movable type is a magnificent way of realizing good designs, and photocomposition already serves equally well. The typewriter is a mediocre way, but far superior to the commercially standard high-speed computer printer as it is often used by novices in automatic language processing. Typesetting and photocomposition machines can be controlled by computer output; text acquired in manuscript, with the functions of its parts marked, can be converted into typesetting control tape by programs that accept a designer's instructions.

If the entire manuscript flow for a journal passes through a computer, with clear coding of all functional categories, the kinds of information that belong in annual indexes of the journal itself, or in abstract bulletins, or secondary publications of whatever kind, can be tapped off and sent to the places where they will be used. Since coding is not by typography but by function, the secondary publications can be designed independently.

The linguistic level of documentation systems is low. The operations that I have discussed are not complex, although they are sometimes intricate. For efficiency, standards are needed. Representing the variety of characters that appear in the scientific publications of the world, using 6-bit or 8-bit characters, is easy enough, but interchange (among journals, for example) is facilitated if everyone adopts the same good plan. The highest level of linguistic sophistication is reached in the determination of index vocabularies. What terms are to be used, with what significance; and how their applicability to a given document can be computed from its text; these are truly difficult questions, to which only partial answers have been given.

Content analysis is a necessary tool in many areas of psychology--and also in sociology, history, anthropology, etc. When a psychiatrist responds to a patient, he has made a kind of content analysis of the foregoing conversation and concluded that a certain response is therapeutically appropriate. Interviews and personal documents provide the richest source of information about many aspects of man's inner life. Only methods that must be described as subjective and impressionistic, on the one hand, or superficial on the other, have been usable in the absence of programs for language processing.

As other speakers in this seminar are saying, the deepest meaning of a text is reflected by its surface structure in a complex fashion. The content of a text is not its vocabulary, since the units of meaning are not words. The superficial grammatical relationships that appear in a text can best be described as the author's attempt, given the lexical and grammatical resources of his language, to show his audience in a compact and perhaps elegant manner the intricacies of an ideational structure that could be formulated in many other ways. The parts would, in fact, probably have to be expressed differently if they were put in a different context--languages are like that.

The content of a text is a thread running through the deepest meanings of its sentences and paragraphs. A casual reference here becomes the first member of an ironic contrast there. Shifts of topic are governed by rhymes and assonances, by puns, by meaningful relationships, and by emotional similarities and contrasts, to mention only a few possibilities. But even these few are enough to show that the psychologist who would use content analysis seriously must have all the help linguistics can give him. Only the psychologist can say what the affective and cognitive properties of linguistic items are for his human subjects. To be sure he is measuring the properties of the entities his subjects are using, he needs the linguist's help in determining phonological, grammatical, and semantic units and structures. And, in view of the size of even an hour's transcript, the linguist's help must be mediated by a computer.

Linguistic processing of natural-language text encompasses acquisition and storage of files of text, consultation of dictionaries, parsing to reveal the structures by which a given grammar accounts for a given text, and recognition of semantic structures. By these processes, a text can be made ready for analysis in terms of psychological variables. The same processes bring text into a form approximately suited to storage in systems that are already being designed to answer factual questions.

To serve in these applications, linguistics must grow. Semantic theories fall far short of what is needed, and until they are stronger we are not quite certain what semantic data we should be collecting. Syntactic theory has developed remarkably in the last decade; facts about grammar are being collected.

Linguistics must be carried out on a distressingly large scale, however. To process text with a small dictionary and a crude grammar is a little like trying to deliver the mail using a sample survey instead of a census as the address book. Most of the mail goes undelivered. Many researchers in this country will immediately start using a dictionary of English in their work, as soon as one is produced and supplied with a program that will match it against text. A parsing system would be almost as much used. The making of such dictionaries and grammars--they are needed also for other languages--is not exactly research, since much of the knowledge that would go into them is either in the literature or of a fairly trivial kind. Nor is it exactly development, since research skills are needed at every step. It is a stepchild, and an expensive one.

Happily for us all, one application of the computer as a language processor has grown up and promises to pay for the rest. Much of documentation, beginning with typesetting, uses no more of linguistics than now exists, and is profitable. Henceforth the pressure for more knowledge about natural languages and better systems for natural-language processing will almost inevitably be unremitting. The only danger is that this pressure from a single source not have a wholly salutary effect.